REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
6/20/2013	FINAL	May 2009 – Mar 2013

4. TITLE AND SUBTITLE

NEW THEORY AND ALGORITHMS FOR SCALABLE DATA FUSION

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-09-1-0500

5c. PROGRAM ELEMENT NUMBER

5. AUTHOR(S)

DeVore, Ronald, A.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Texas A&M University
Sponsored Research Services
400 Harvey Mitchell Pkwy, Suite 300
College Station, TX 77845

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

USAF, AFRL DUNS 143574726

Air Force Office of Scientific Research
875 N. Randolph St.
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-OSR-VA-TR-2013-0368

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION A: APPROVED FOR PUBLIC RELEASE

13. SUPPLEMENTARY NOTES

14. ABSTRACT

New mathematical theory and algorithms were developed for querying functions and data sets in high dimensions. This was accomplished in both the deterministic setting as well as the stochastic setting with noise. Since high dimensional problems suffer from the curse of dimensionality, new model classes were introduced based on the notions of sparsity and variable reductions. These model classes were shown to fit real world problems and yet be amenable to realistic computational methods for querying in high dimensions. The new algorithms were developed using sophisticated adaptive methods. They utilize the most judicious deterministic point clouds in high dimension such as those based on discrepancy theory (Halton sequences) and perfect hashing. The resulting algorithms were shown to have optimal performance on the proposed model classes. Thus, they have provided the most efficient algorithms for querying high dimensional data under the model assumptions. Results were also developed and proved for classification of data. These stochastic algorithms were shown to have optimal performance on various model classes.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE		5	Dr. Roanld A. DeVore
					19b. TELEPHONE NUMBER (Include area code) 979-845-3336

To: technicalreportsafosr.af.mil
Subject: **Annual Progress Statement to Dr. Douglas Cochran**
Grant Title: NEW THEORY AND ALGORITHMS FOR SCALABLE DATA FUSION
Grant Number: FA9550-09-1-0500
Reporting Period: 6/15/2009 to 03/31/2013

To: technicalreportsafosr.af.mil
Subject: **Final Progress Report to Dr. Tristan Nguyen**
Grant Title: NEW THEORY AND ALGORITHMS FOR SCALABLE DATA FUSION
Grant Number: FA9550-09-1-0500
Reporting Period: Final Report

Abstract for SF 298: New mathematical theory and algorithms were developed for querying functions and data sets in high dimensions. This was accomplished in both the deterministic setting as well as the stochastic setting with noise. Since high dimensional problems suffer from the curse of dimensionality, new model classes were introduced based on the notions of sparsity and variable reductions. These model classes were shown to fit real world problems and yet be amenable to realistic computational methods for querying in high dimensions. The new algorithms were developed using sophisticated adaptive methods. They utilize the most judicious deterministic point clouds in high dimension such as those based on discrepancy theory (Halton sequences) and perfect hashing. The resulting algorithms were shown to have optimal performance on the proposed model classes. Thus, they have provided the most efficient algorithms for querying high dimensional data under the model assumptions. Results were also developed and proved for classification of data. These stochastic algorithms were shown to have optimal performance on various model classes.

Final Report: Modern data processing is challenged by the size and variety of data that must be queried to extract the relevant details needed to answer targeted questions. Mathematically, the problem can be formulated as learning a function depending on a large number of variables. The type of learning problem encountered depends on the information available about the function. In *directed learning* we are allowed to query the function at any points we wish but at perhaps a very high cost and the question is where should we direct our queries. In *stochastic learning*, the values are given to us as random draws (usually with noise) and the question is how to best utilize this information.

The inherent impediment to learning in high dimension is the so called *curse of dimensionality* which says our ability to recover a smooth function deteriorates significantly with increasing dimension. The only way around this is if the target functions satisfies additional properties which reduce their complexity. The problem is to identify mathematically what these properties may be and then utilize them in the development of learning and computation algorithms. The goal of the research project is to develop new

model classes which represent real world problems in high dimension, yet do not suffer the curse of dimensionality. These models are based on the view that only a small number of the variables are significant and if these can be found then a model reduction can be performed to make the learning problem more tractable. Its main focus is to develop new model classes in high dimension that give a precise description of this variable reduction and then to develop algorithms that perform at optimal distortion in terms of the computational budget for each of these new classes. The PI and his collaborators have put forward several model classes for functions of many variables and shown how to develop optimal query algorithms for each of these model classes. In addition these models use sparsity and multiscale expansions to further increase their tractability. The following is a summary of the notable achievements under this research project.

Directed learning in high dimensions With Guergana Petrova and Przemek Wojtaszczyk, the PI has investigated the problem of capturing a continuous function f on a compact domain $\Omega \subset I\!R^N$ with N large by querying the point values of f. This problem is sometimes called *directed learning*. It is known that this problem suffers from the curse of dimensionality which means that even smooth functions need an impossible number of queries. For example, for a general function in C^s, we would need $\epsilon^{-N/s}$ queries to capture it to accuracy ϵ. The only way to defeat this curse and have reasonable numerical algorithms is to assume more about f. The dominant theme in applications is to assume that f actually depends on much fewer variables (unknown to us) or it can be well approximated by functions of few variables. In our work [10], we have shown that under both of these models we can break the curse of dimensionality by tailoring our questions. This work is very foundational and of interest also in computer science and statistics.

The work described above has captured the interest of several other researchers. With Albert Cohen , Ingrid Daubechies, Gerard Kerkyacharian, and Dominique Picard [6], the PI has extended the results of [10] to include other models for functions. The simplest examples are the ridge functions $f(x) = g(a \cdot x)$. We show how to capture such a function f with g and a unknown by querying its point values at specially constructed point sets. Our approach is a combination of hashing and compressed sensing techniques. Our first versions of these algorithms were not stable in terms of perturbations of the queries. We have recently uncovered how to handle the instability.

Stochastic learning: We revisited the fundamental problem of classification from randomly drawn data with a particular eye towards high dimensional problems [3]. We have developed new ways to bound variance in terms of VC dimension. We also developed new model classes for the regression function which reflect properties which enable fast algorithms even in high dimension. One difficulty in treating high dimensional problems is that standard localization techniques such as adaptive partitioning are impossible to implement in their standard form because one adaptive refinement results in an inordinate number of cells. As an alternative, we have developed with Peter Binev and Wolfgang Dahmen a new theory for adaptive partitioning called *sparse occupancy trees*. The cardinality of these trees is controlled by the initial size of the data. Therefore they become numerically feasible since the number of computations is linear in the size of the data.

This has allowed us to develop learning algorithms in [5] for regression problems in high dimension which have several advantages over the existing techniques based on nearest neighbors or support vector machines. A follow up giving an analysis of the numerical performance of these new classification algorithms is given in [4]

Querying manifolds in high dimension: One of the common ways of avoiding the curse of dimensionality is to assume that the data comes from a smooth low dimensional manifold in high dimensions. We have developed greedy algorithms for taking snapshots of these manifolds and then using them to create good linear space fits to the manifold [2]. We have been able to accomplish this even in the setting of infinite dimensional Banach spaces [11]. These techniques are heavily used in model reduction for PDEs.

Stochastic and parametric PDEs: Such PDEs model many physical systems. The standard approach to solving stochastic PDEs are Monte Carlo methods. In [7], we offer an attractive alternative to Monte Carlo for solving elliptic stochastic equations. Our methods are based on Wiener chaos expansions which convert the stochastic equations to parametric equations with an infinite number of parameters. We show that the resulting parametric equations have a sparse representation with respect to polynomial basis in the parameters. This lays the foundation for efficient numerical methods to solve the original stochastic equation. They solve infinite dimensional parametric problems in polynomial time.

This original work gave dramatic improvement in computational efficiency over Monte Carlo. However, some of the assumptions on the diffusion coefficients were not natural. In [8], we introduce complex variable methods to obtain what seem to be the most natural and far reaching results in terms of efficiency of recovery.

This work on stochastic PDEs has been implemented in numerical algorithms but not to the full extent of the theory. With Chkifa, Cohen and Schwab [9], we have given a greedy procedure which will generate optimal Wiener chaos polynomial spaces in which to capture the solution to the stochastic PDE. These algorithms are an analogue of the famous wavelet greedy algorithms developed by Cohen-Dahmen-DeVore. They display dramatic improvement in their rate distortion over all existing algorithms.

Tensor structure: One of the common themes to handle high dimensional structure in quantam chemistry is to assume a tensor structure. There are many possible tensor formats and a coherent theory for approximation by tensors has not yet been developed. We have given first steps toward such a theory in [1] where querying algorithms for low rank tensors are considered.

References

[1] M. Bachmayr, W. Dahmen, R. DeVore, and L. Grasedyck, *Approximation of High-Dimensional Rank One Tensors*, Constructive Approximation, to appear.

[2] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk, *Convergence rates for greedy algorithms in reduced basis methods*, SIAM Journal of Math. Analysis, **43**(2011), 1457–1472

[3] P. Binev, A. Cohen, W. Dahmen, R. DeVore, *Classification algorithms using adaptive partitioning*, submitted.

[4] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and P. Lamby, *An analysis of tree based algorithms for classification*, in preparation

[5] P. Binev, W. Dahmen, R. DeVore, and P. Lamby, *Sparse occupancy trees and learning in high dimensions*, in preparation.

[6] A. Cohen, I. Daubechies, R. DeVore, G. Kerkyacharian and D. Picard *Capturing Ridge Functions in High Dimensions from Point Queries*, Constructive Approximation, **35**(2012), 225–243.

[7] A. Cohen, R. DeVore, and C. Schwab, *Convergence rates of best Galerkin approximation for a class of elliptic sPDEs*, J. FOCM, **6**(2010), 615–646.

[8] A. Cohen, R. DeVore, and C. Schwab, *Analytic regularity for parametric and stochastic elliptic PDEs*, Analysis and Applications, **9**(2011), 11–47

[9] A. Chkifa, A. Cohen, R. DeVore, and C. Schwab, *Adaptive algorithms for sparse polynomial approximation of parametric and stochastic elliptic PDEs*, M2AN Math. Model. Numer. Anal., **47**(2013), 253–280.

[10] R. DeVore, G. Petrova, and Przemyslaw Wojtaszczyk, *Approximation of functions of few variables in high dimensions*, Constructive Approximation, **33**(2011),125–143.

[11] R. DeVore, G. Petrova, and Przemyslaw Wojtaszczyk, *Greedy Algorithms for Reduced Bases in Banach Spaces*, Constructive Approximation, to appear.